REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

AD-A265 102

...o average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and ... on of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including ... Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA ... Reduction Project (0704-0188), Washington, DC 20503

2. REPORT DATE
May 1993

3. REPORT TYPE AND DATES COVERED
professional paper

4. TITLE AND SUBTITLE
PERFORMANCE MEASURES FOR NEURAL NETS USING JOHNSON DISTRIBUTIONS

5. FUNDING NUMBERS
PR: ST56
PE: 0603741D
WU: DN302231

6. AUTHOR(S)
W. C. Torrez, J. T. Durham, R. D. Trueblood

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Command, Control and Ocean Surveillance Center (NCCOSC)
RDT&E Division
San Diego, CA 92152-5001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Space and Naval Warfare Systems Command
Washington, DC 20363-5100

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

A probability distribution for multilayer perceptron artificial neural net outputs is derived assuming a sigmoidal activation function. This distribution is known to be a member of the Johnson system of distributions. Using this distribution, theoretical receiver operating characteristic curves can be developed to obtain recognition differential values for corresponding values of the probability of false alarm. Application of these techniques for the detection of broadband signals is presented.

Published in *Proceedings of the International Conference on Neural Networks*, March 1993, Vol. 1, pp. 506-510.

14. SUBJECT TERMS
neural networks
active signal processing

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAME AS REPORT

UNCLASSIFIED

21a. NAME OF RESPONSIBLE INDIVIDUAL	21b. TELEPHONE (Include Area Code)	21c. OFFICE SYMBOL
W. C. Torrez	(619) 553-2020	Code 733

DTIC QUALITY INSPECTED 6

Accesion For	
NTIS CRA&I	✓
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	20

NSN 7540-01-280-5500

Standard form 298 (BACK)

UNCLASSIFIED

Proceedings of the IEEE International Conference on Neural Networks, Vol. 1, pp. 506-510, March 1993.

Performance Measures for Neural Nets Using Johnson Distributions

William C. Torrez and Jayson T. Durham
Signal and Information Processing Division
NCCOSC RDT&E Division

Richard D. Trueblood
ORINCON Corporation
San Diego, CA

***Abstract*— A probability distribution for multilayer perceptron artificial neural net outputs is derived assuming a sigmoidal activation function. This distribution is known to be a member of the Johnson system of distributions. Using this distribution, theoretical receiver operating characteristic curves can be developed to obtain recognition differential values for corresponding values of the probability of false alarm. Application of these techniques for the detection of broadband signals is presented.**

I. Introduction

In this paper, we consider a feedforward multilayer perceptron trained with back propagation of error. The output nodes in one layer are transmitted to nodes in another layer through links that amplify or attenuate such outputs through weighting factors. Except for the input layer nodes, the net input to each node is the sum of the weighted outputs of the nodes in the prior layer. Each node is activated in accordance with the input and bias to the node, and the activation function of the node. The typical activation function for the nodes in the hidden layers is the common sigmoid or logistic function,

$$f(a) = \frac{1}{1+e^{-(a-\theta)/\theta_o}}. \tag{1}$$

This function is also known, in neural net terminology, as the squashing function. In (1), the parameter θ serves as a threshold or bias and the parameter θ_o modifies the shape of the sigmoid. It is the objective of this paper to evaluate the performance of the neural net by modeling the class conditional probability density functions $p_{c_i|\mathbf{x}}, i = 0, 1$, for noise alone and for signal plus noise, respectively, using the sigmoidal squashing function. Although the detection and false alarm statistics are unchanged, it will be seen that the bias and shape parameters characterize these distributions. Here $\mathbf{x}$ is the input feature vector to be classified, and the output class c_0, represents the noise alone case, while class c_1 represents the signal plus noise case. The performance metric presented in this paper is based on the class conditional pdf's and is known as the receiver operating characteristic (ROC) curve, which presents the probability of detection $P_d = \int_t^\infty p_{c_1|\mathbf{x}}(\tau)d\tau$ as a function of detection threshold t, where t is chosen to achieve some prescribed level of probability of false alarm, $P_{fa} = \int_t^\infty p_{c_0|\mathbf{x}}(\tau)d\tau$.

In II, it will be shown that under certain conditions the pre-sigmoided hidden layer input is an approximate normal random variable. It follows that, in these cases, the sigmoided output is a logistic transformation of an approximate normal random variable. In III, the distribution of this transformation will be derived and will be identified as a member of the Johnson system of distributions. Using this model identification, ROC curves can be plotted as a function of detection threshold t and parameterized by signal-to-noise ratio (SNR). Another metric of practical utility is the recognition differential (RD), which is the SNR which guarantees probability of detection equal to 1/2 for a prescribed level of false alarm probability; in practical applications, this may in fact be the preferred metric. This point will be made in IV for the detection of certain broad-band transient signals generated by simulation in our laboratory.

II. Sums as Gaussian Distributions

We consider now a multilayer perceptron with M continuous valued inputs $\mathbf{x} = (x_1, \ldots, x_M), x_k \epsilon (-\infty, \infty)$, and two layers of hidden nodes. Taking the Bayesian approach allows us to consider x_k as a random variable; we will assume that its mean μ_k and variance σ_k^2 are finite. Without loss of generality, we assume that $\mu_k = 0$. The net input

to hidden node j may be expressed as

$$X_j = \sum_{k=1}^{M} w_{kj} x_k,$$

for real-valued weights, $\{w_{kj}\}$.

The requirement that the terms $w_{kj}x_k$ be uniformly small, which is known as the Lindeberg condition [1], is a sufficient condition to insure that the sums X_j, properly normalized, converge to the normal distribution. We have made the empirical observation in the laboratory that, in certain cases, the independent random variables $w_{kj}x_k$ do, in fact, have a uniformly small effect on the sum X_j. Thus, our contention that, in these cases, the sums X_j follow a normal distribution is borne out by the application of the central limit theorem as described above.

For the present application, M is sufficiently large (on the order of 1440) and the x_k are sufficiently decorrelated to insure a high degree of statistical independence in the collected samples, so we may, in fact, invoke the central limit theorem to assert normality when the Lindeberg condition holds. To insure independence, if x_k is a portion of continuous time series, then we assume that x_k has been sampled at a rate which is higher than the decorrelation time of the time series. We could also pre-whiten the time series by Gram-Schmidt orthogonalization techniques to insure a high degree of independence.

III. Sigmoidal Squashing Function and the Johnson Distributions

Recall the squashing function, $f(a)$, defined in (1),

$$f(a) = \frac{1}{1 + e^{-(a-\theta)/\theta_o}}, -\infty < a < \infty.$$

Assuming that X_j has an approximate normal distribution with mean 0 and variance $\overline{\sigma}_j{}^2 = \sum_{k=1}^{M} w_{kj}{}^2 \sigma_k{}^2$ as asserted in the previous section, the probability density function, $p_{c|\mathbf{x}}(s)$, of $f(X_j)$ can be easily derived. In fact by the change of variables formula,

$$p_{c|\mathbf{x}}(s) = \phi\left(\theta + \theta_o \ln \frac{s}{1-s}\right) \frac{dz}{ds}, 0 < s < 1, \qquad (2)$$

where $\phi(z) = \frac{1}{\overline{\sigma}_j \sqrt{2\pi}} e^{-z^2/2\overline{\sigma}_j{}^2}, -\infty < z < \infty$ and $z = \theta + \theta_o \ln\left(\frac{s}{1-s}\right)$. After differentiating z with respect to s in (2), we obtain

$$p_{c|\mathbf{x}}(s) = \frac{\eta_j}{\sqrt{2\pi}} \frac{1}{s(1-s)} e^{-\frac{1}{2}\left\{\gamma_j + \eta_j \ln \frac{s}{1-s}\right\}^2}, 0 < s < 1, \qquad (3)$$

where, $\eta_j = \theta_o/\overline{\sigma}_j, \gamma_j = \theta/\overline{\sigma}_j$. The density in (3) is a member of the Johnson system of distributions and its properties are well-known [2]. In fact, maximum likelihood estimates of γ_j and η_j, are given by

$$\hat{\gamma}_j = -\overline{X}_j/s_j, \hat{\eta}_j = 1/s_j, \qquad (4)$$

where $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ji}$ and $s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(X_{ji} - \overline{X}_j)^2$, where $\{X_{ji} : i = 1, \ldots, n\}$ are sampled from the hidden layer at node j.

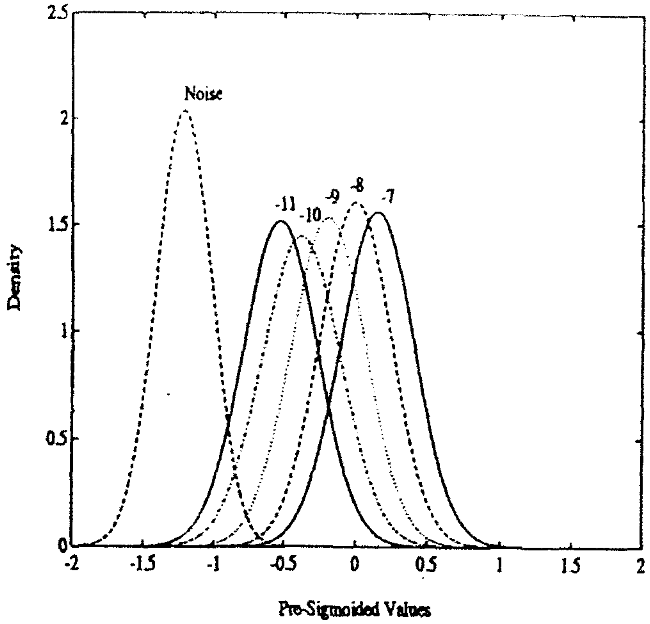

Figure 1: Fitted Normal Density Plots for Pre-Sigmoided Values Indexed by Signal Level Offsets (in dB) Including the Case of Noise Only.

IV. Receiver Operating Characteristic Curves for Broad-Band Signals

The pre-sigmoided values were collected from the data sets described in the Appendix for the five cases of signal mixed with noise at various levels offset in 1 dB increments from a reference signal and for the noise alone case. Fitted normal density plots for these values are shown in Fig. 1. In each of these cases, the observed pre-sigmoided values passed the Kolmogorov-Smirnov (K-S) and the chi-square goodness of fit tests for normality at the 5 % level of significance as we had asserted in II.

Fig. 2 shows the corresponding fitted Johnson density plots, where the form of the pdf is given by (3). The fits were based on the parameter estimates, $\hat{\gamma}_j$ and $\hat{\eta}_j$ given in (4). The signal-to-noise ratio (SNR) in dB offset is noted on each density curve. The SNR values given have been calculated for the simulated broad-band transient signals according to recent work described in [3]. The reader is

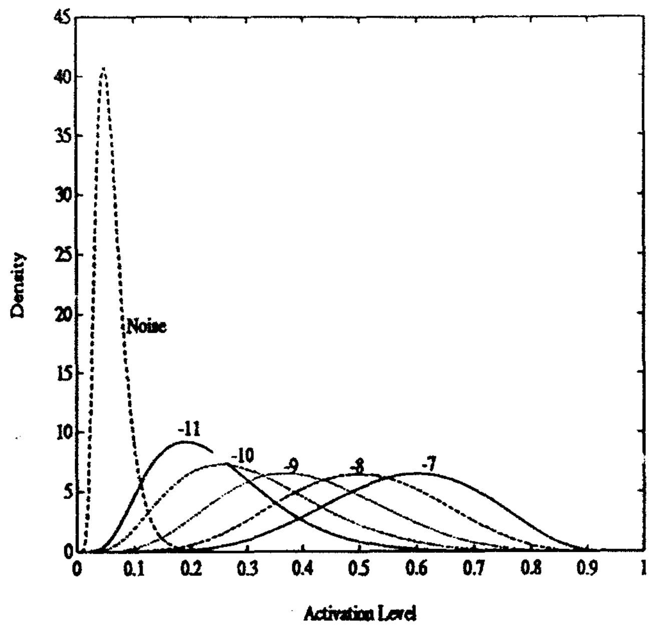

Figure 2: Fitted Johnson Density Plots for Sigmoidal Outputs Indexed by Signal Level Offsets (in dB) Including the Case of Noise Only.

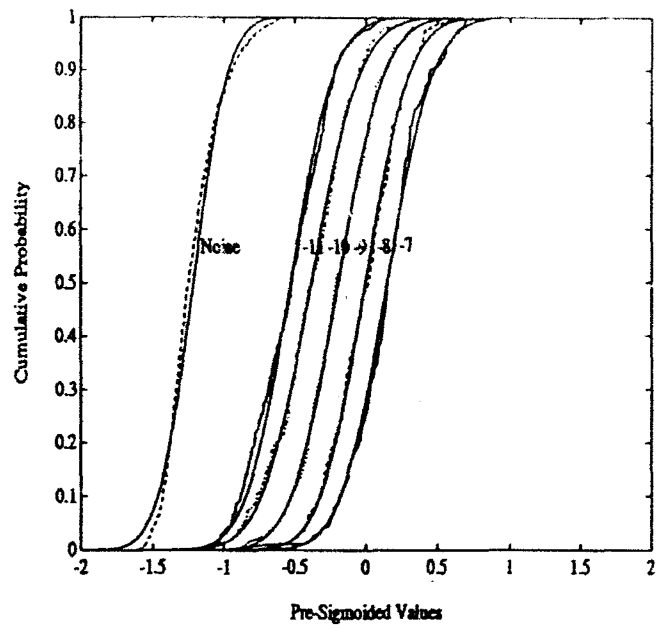

Figure 3: Empirical Distribution Functions for Pre-Sigmoided Values Overplotted with Semi-empirical Normal Cumulative Distribution Functions.

referred to that source for the technical details for the SNR calculations.

Fig. 3 shows the fitted normal cumulative distribution functions overplotted with the empirical distribution function for the pre-sigmoided values. Fig. 4 shows the corresponding semi-empirical fits of the Johnson cumulative distribution functions (smooth curves) to the empirical distribution functions of the sigmoided outputs (i.e., activation levels) for the five levels of signal power considered as well as the noise alone case. These semi-empirical fits using the Johnson distribution were deemed statistically close to the empirical observations as measured by the K-S goodness of fit test performed at the 5 % level of significance.

Finally Fig. 5 gives the ROC curves for the various signal level offsets based on the semi-empirical (i.e., fitted) Johnson distributions. These plots show P_d as a function of P_{fa}. One sees that for a prescribed level of P_{fa} of 10^{-5} and $P_d = \frac{1}{2}$, the signal level offset is very close to -10 dB. This value is known as the recognition differential offset or RD offset (for the prescribed level of P_{fa}).

Also Fig. 6 below shows a plot of the means and standard deviations of the five signal levels considered as a function of the level offset. Beyond allowing a straightforward interpolation, the fitted least squares line also allows us to extrapolate the mean and standard deviation of non-

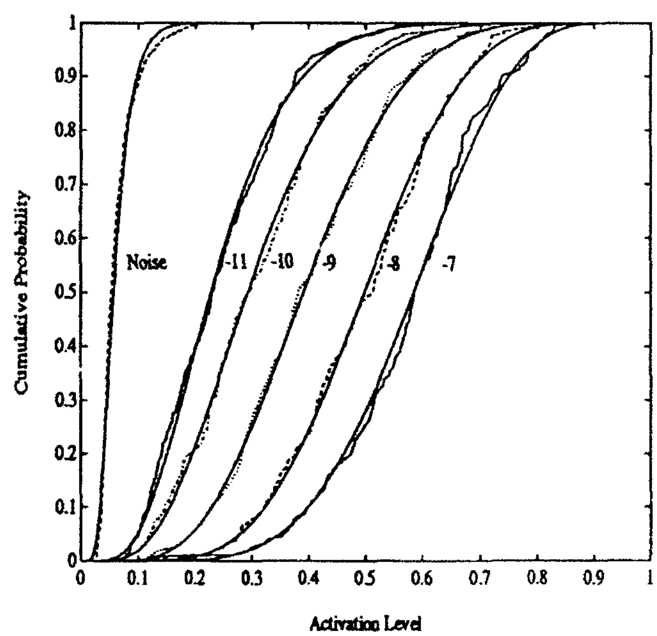

Figure 4: Empirical Distribution Functions for the Sigmoidal Outputs Overplotted with Semi-empirical Johnson Cumulative Distribution Functions.

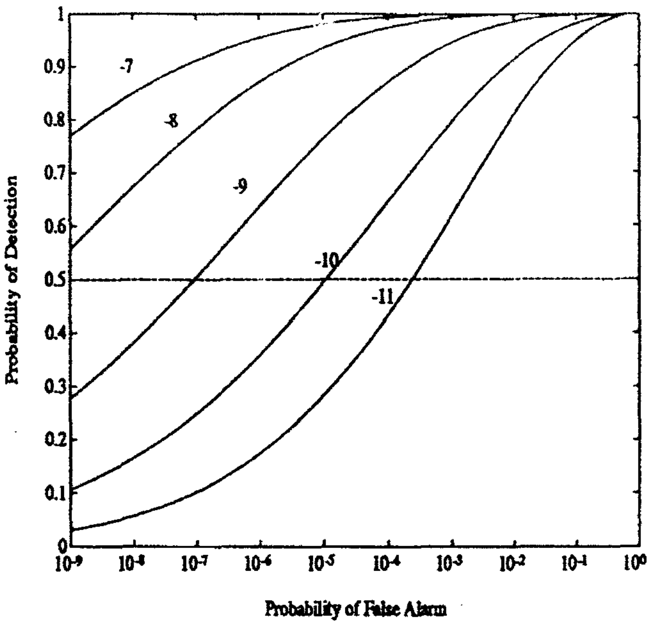

Figure 5: Semi-empirical Receiver Operating Characteristic Curves Using Johnson Distributions for the Sigmoided Outputs Indexed by Signal Level Offsets (in dB).

simulated signal levels. The equations for the mean and standard deviation fits are respectively,

$$y = 1.3722 + 0.17381x, \tag{5}$$

$$y = 0.22241 - 0.004246x. \tag{6}$$

Now to obtain the RD given any prescribed level of P_{fa}, we observe that the detection threshold, $t_d(\alpha)$, is given by

$$t_d(\alpha) = s_o\Phi^{-1}(1-\alpha) + m_o, \tag{7}$$

where $m_o = -1.21474$ and $s_o = 0.196038$ are the mean and standard deviation,respectively, of the noise, and Φ^{-1} is the inverse of the unit normal cumulative distribution function. Using (5), it is easy to see that the RD at level α must satisfy

$$RD = \frac{t_d(\alpha) - m_1}{m_2},$$

where $m_1 = 1.3722$ and $m_2 = 0.17381$. Together with (7), we have finally,

$$RD = \frac{s_o\Phi^{-1}(1-\alpha) + m_o - m_1}{m_2} \tag{8}$$

Finally, Fig. 7 plots RD as function of P_{fa} given by (8).

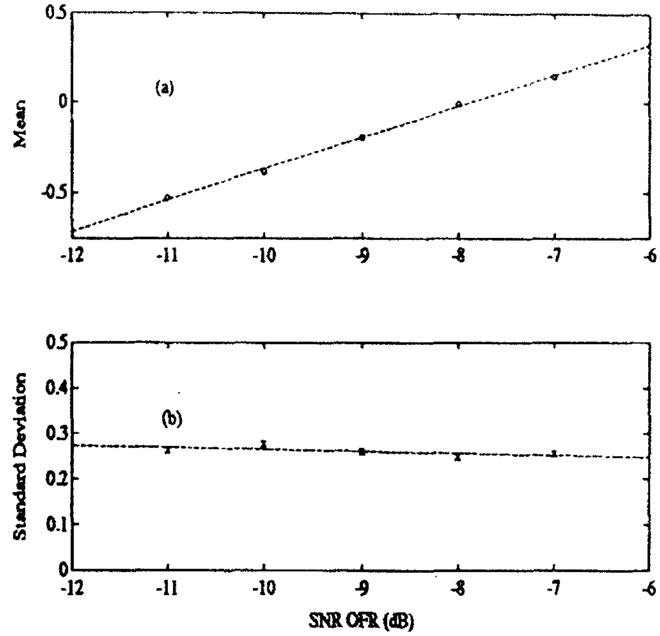

Figure 6: (a) Means and (b) Standard Deviations of Five Cases with Linear Fits.

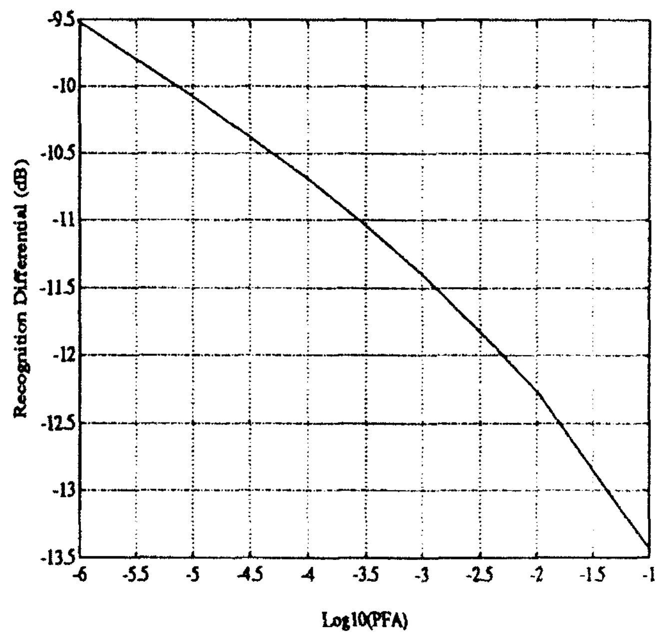

Figure 7: Recognition Differential as a Function of P_{fa}.

V Conclusions

For a prescribed level of P_{fa} of 10^{-3}, the RD may be extrapolated by noting a functional relationship of the means and variances of the normal distributions of the pre-sigmoided values as a function of the signal level offset (cf. Fig. 1). Extrapolating this value gives an RD of approximately -11.4 dB (cf. Fig. 7) offset from the reference signal.

As mentioned in II, the invocation of the central limit theorem for establishing the normality of the weighted sums of sigmoided neuron outputs is only applicable when the Lindeberg condition holds. Our most recent work [4] demonstrates that we can, in fact, fit the Johnson system of distributions to empirical distributions of various shapes and that, like the results described in this paper, these generalized fits can also describe the location and shape of the distributions in terms of the input SNR levels.

Acknowledgment

We wish to thank Mr. Lou Griffith, Project Manager at NRaD for his support and encouragement during the preparation of this report.

References

1. B. V. Gnedenko, *The Theory of Probability*, 4th ed., New York, N. Y.: Chelsea, 1968, p. 305.
2. N. L. Johnson, "Systems of frequency curves generated by methods of translation," *Biometrika*, vol. 36, 1949, pp. 149-176.
3. J. T. Durham, W. C. Torrez, and E. W. VonColln, "Performance analysis of the Air Defense Initiative neural network processor using data sets from the E1 test," *NRaD Tech. Document*, in press.
4. W. C. Torrez and J. T. Durham, "Johnson distributions for fitting weighted sums of sigmoided neuron outputs," unpublished.

Appendix Experimental Methodology

A strong SNR broadband transient event was digitized from a recorded data tape. A 30 second interval, which contained the signal with a trailer of noise, was then repeatedly played to a tape recorder for two hours. There were no gaps between the beginning and ending noise samples of the captured interval. Edge effects were an initial concern but none have been observed. The recorded tape of repeated events was designated as the signal master tape. Using an analog signal attenuator, bandpass filtered output from the signal master was then recorded to a series of tapes with the analog attenuator offset one dB from the reference signal, per recorded tape. The initial tape, which was consequently recorded at the strongest signal level, was designated as the reference signal level. Each two hour tape stored 240 repetitions of the same event.

Finally, all the signal tapes were analog mixed with a two hour period of ambient ocean noise. For a given time on the noise tape, the time of occurrence of the events varied up to within a few seconds. None of the events were mixed with the noise to within the same sampling interval. By not playing a signal tape, a noise only tape was recorded. For all the test recordings, the two-hour noise interval was simply repeated for each signal tape with all other system parameters fixed.

Each tape was played into a realtime classification system which utilizes in-house developed artificial neural networks. For noise only, the output activation levels were blocked into contiguous signal length durations and the maximum activation level was recorded for each block. Those values were used to calculate the empirical distribution of noise only. For each signal tape, the maximum activation plus or minus the signal duration was then recorded. Those values were used to calculate the signal present distributions.